哈尔滨工业大学（深圳）

大数据导论大作业报告

题目：特定疾病的回归和分类

姓　　名　　胡聪

学　　号　　180110505

报告日期　　2021.1.7

一、 实验目的

本实验旨在通过某种患病病人的临床数据和体检指标来预测人群指示病情程度的指标。需要设计高效，且解释性强的算法来精准预测病情指标。全部编程实现。

二、 实验内容分析

实验阶段Ⅰ数据为训练文件 d_train.csv。每个文件第一行是字段名，之后每一行代表一个个体。文件共包含 42 个字段，包含数值型、字符型、日期型等众多数据类型，部分字段内容在部分人群中有缺失，其中第一列为个体 ID 号。训练文件的最后一列为标签列，既需要预测的目标值。提交说明：提交一个 d_model.py 预测的模型文件。实验阶段Ⅱ数据为训练文件 f_train.csv。每个文件第一行是字段名，之后每一行代表一个个体，部分字段名已经做脱敏处理。文件共包含 85 个字段，部分字段内容在部分人群中有缺失，其中第一列为个体 ID 号。训练文件的最后一列为标签列，既需要预测的是否患病的类标。提交说明：提交一个 f_model.py 文件。

实验的总体流程分为数据可视化、数据预处理、特征工程、模型融合等。数据可视化可以验证对于实验数据分布的一些猜想，让我们对数据分布有一个清晰的认识和理解，对数据预处理环节有很大的帮助。数据清洗中需要检测异常样本、对缺省字段进行处理，还要进行数据采样，在数据正负样本不均衡的情况下，准确率很好，但是在测试集效果不佳，泛化能力弱，因此需要使得数据样本均衡。在特征工程中，可以把数据大概分为数值型、类别型、时间型、文本型、统计型、组合特征等，这里可以使用 sklearn 进行处理。还需要进行连续特征离散化，起到简化逻辑回归模型的作用，降低模型过拟合的风险。然后通过查看当前参数的当前模型过拟合状态还是欠拟合状态。最后选择算法进行模型融合。

三、 实验过程及结果

利用 pandas 读取数据，因为数据中含有中文，所以需要采用 gbk 编码进行读取。

```
df = pd.read_csv("./d_train.csv",encoding="gbk")
df = pd.read_csv("./f_train.csv",encoding="gbk")
```

d_train shape (5642, 42)

f_train shape (1000, 41)

可以读到，训练集有 5462 条数据，42 列，而我们需要预测的数据有 1000 条，41 列（训练集中含有血糖数据，多一列）。

print(d_train.columns)

```
Index(['id', '性别', '年龄', '体检日期', '*天门冬氨酸氨基转换酶', '*丙氨酸氨基转换酶', '*碱性磷酸酶',
       '*r-谷氨酰基转换酶', '*总蛋白', '白蛋白', '*球蛋白', '白球比例', '甘油三酯', '总胆固醇',
       '高密度脂蛋白胆固醇', '低密度脂蛋白胆固醇', '尿素', '肌酐', '尿酸', '乙肝表面抗原', '乙肝表面抗体', '乙肝 e 抗原',
       '乙肝 e 抗体', '乙肝核心抗体', '白细胞计数', '红细胞计数', '血红蛋白', '红细胞压积', '红细胞平均体积',
       '红细胞平均血红蛋白量', '红细胞平均血红蛋白浓度', '红细胞体积分布宽度', '血小板计数', '血小板平均体积',
       '血小板体积分布宽度', '血小板比积', '中性粒细胞%', '淋巴细胞%', '单核细胞%', '嗜酸细胞%', '嗜碱细胞%',
       '血糖'],
      dtype='object')
```

首先我们需要了解数据集中的有关信息，通过分析数据集我们可以发现乙肝表面抗原、乙肝表面抗体、乙肝核心抗体、乙肝 e 抗原、乙肝 e 抗体的缺失较多，比例很大，因此初步想法是删除相关数据。

然后分析数据影响权重，可以分析数据集发现，样本的体检日期与性别对于模型的预测基本没有关联，即样本的血糖值与体检日期和性别无关联，因此可以删除相关数据内容，去除对模型无影响的基本特征和离群值。

对于缺失值，我们应该对数据值中缺失的数据采取恰当的方式进行填充处理。对于缺失值，有多种处理方法，例如特殊值填充、平均值填充、中位数填充、回归方程、期望值最大化等，这里个人使用的是中位数填充的方法进行缺失值数据处理。

特征平滑处理这一方面，并非所有的基本特征都需要进行平滑处理，比如说淋巴细胞就不需要进行特征平滑处理。

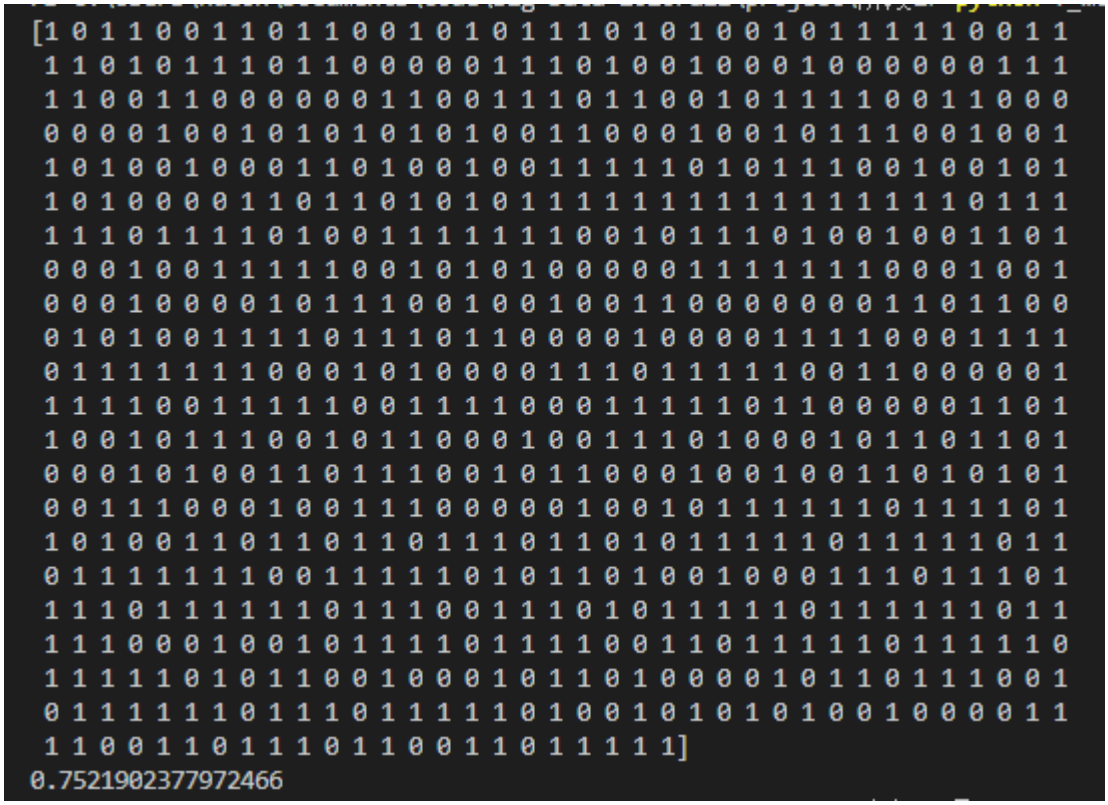
数据特征工程采用 sklearn 中的 PolynomialFeature 产生相互影响的特征集，采用线性回归的方式来做非线性回归的预测。

实验结果如下：

阶段 1

```
<class 'numpy.ndarray'>
平均绝对值误差 0.0029291154071470417
均方差 0.0029291154071470417
中值绝对值误差 0.0
PS C:\Users\hucon\Documents\Code\Big-Data-2020Fall\project\阶段1>
```

阶段 2



四、 实验心得

本次大数据实验的大体流程可以分为数据可视化、数据预处理、特征工程和模型融合。在数据预处理过程中，有很多坑需要注意，我们需要去除异常样本，还需要对缺省字段进行处理，例如有缺省值较多、连续特征缺省值处理、非连续特征缺省处理等，而对于缺省较少的情况下，可以考虑多种填充方法。总的来说通过这次大作业收获了很多数据挖掘相关工具的使用方法，以及了解了相关的算法，对机器学习和数据处理有了更加深入的认识和了解。